HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

CƠ SỞ TẠI THÀNH PHỐ HỒ CHÍ MINH

BÁO CÁO CUỐI KÌ

MÔN HỌC: **DATA MINING**

Đề tài: Đánh giá chất lượng nhà cung cấp

Giảng viên hướng dẫn : Nguyễn Ngọc Duy

Sinh viên thực hiện:

N20DCCN135 :Nguyễn Văn Tài
N20DCCN118 : Trang Tuấn Minh
Lớp: D20CQCNHT01-N

TPHCM-2024

Contents

LỜI MỞ ĐẦU

Trong thời đại công nghệ thông tin và mạng lưới internet ngày nay, việc thu thập và phân tích dữ liệu đã trở thành một công cụ quan trọng không chỉ trong việc nắm bắt xu hướng thị trường mà còn trong việc đưa ra các quyết định kinh doanh chiến lược. Đặc biệt, trong lĩnh vực kinh doanh may mặc, việc xây dựng một kho dữ liệu toàn diện về cả hai hình thức kinh doanh online và offline đóng vai trò quan trọng trong việc hỗ trợ các doanh nghiệp trong quá trình ra quyết định chiến lược về hình thức kinh doanh.

Đề tài nghiên cứu "Xây dựng kho dữ liệu về kinh doanh online và offline mặt hàng may mặc và khai phá kho dữ liệu này cho mục đích ra quyết định về hình thức kinh doanh online hoặc offline" được thực hiện nhằm mục đích cung cấp một cái nhìn tổng quan về thị trường may mặc cũng như hỗ trợ các doanh nghiệp trong việc đưa ra quyết định chiến lược về cách thức kinh doanh phù hợp nhất.

Báo cáo này sẽ tập trung vào quá trình xây dựng và khai thác kho dữ liệu về kinh doanh may mặc online, bao gồm cách thức thu thập dữ liệu, phân tích và hiểu biết về xu hướng thị trường, hành vi của người tiêu dùng và các yếu tố ảnh hưởng đến quyết định kinh doanh. Chúng tôi cũng sẽ trình bày một số phương pháp và công cụ phân tích dữ liệu để hỗ trợ việc ra quyết định chiến lược cho các doanh nghiệp trong ngành may mặc.

Trong quá trình hoàn thành đề tài chúng em đã gặp phải một số khó khăn do việc tìm hiểu kỹ thuật khai phá dữ liệu còn mới mẻ, khối lượng kiến thức trong lĩnh vực khai phá dữ liệu còn nhiều và liên tục được cập nhật nên chắc chắn không tránh khỏi những sai sót. Kính mong sự đóng góp ý kiến của GV để chúng em cố gắng hoàn thiện tốt hơn

Nội dung nghiên cứu gồm 5 chương:

CHƯƠNG 1 : Tổng quan về khai phá dữ liệu

CHƯƠNG 2 : Giới thiệu kỹ thuật phân cụm trong khai phá dữ liệu

CHƯƠNG 3: Giới Thiệu Kỹ Thuật Phân Cụm Trong Khai Phá Dữ Liệu

CHƯƠNG 4: Tìm hiểu về công cụ WEKA

CHƯƠNG 5: Triển khai phân cụm dữ và đánh giá kho dữ liệu

CHƯƠNG 1:GIỚI THIỆU ĐỀ TÀI

1.Lý do chọn đề tài

Đánh giá chất lượng nhà cung cấp" xuất phát từ tầm quan trọng của việc tối ưu hóa chuỗi cung ứng và nâng cao hiệu quả hoạt động của doanh nghiệp. Đánh giá chất lượng nhà cung cấp giúp đảm bảo rằng các nhà cung cấp đạt tiêu chuẩn chất lượng cao, từ đó duy trì và nâng cao chất lượng sản phẩm cuối cùng. Việc này còn giúp giảm thiểu rủi ro bằng cách phát hiện và xử lý sớm các vấn đề từ nhà cung cấp kém chất lượng, đảm bảo sự liên tục và ổn định của chuỗi cung ứng. Ngoài ra, các nhà cung cấp chất lượng cao thường hoạt động hiệu quả hơn, giúp giao hàng đúng hạn và giảm lãng phí, từ đó tăng cường hiệu suất chung của doanh nghiệp. Hơn nữa, đánh giá chất lượng nhà cung cấp cũng tạo điều kiện xây dựng mối quan hệ hợp tác bền vững, tạo động lực cho nhà cung cấp duy trì và cải thiện dịch vụ, dẫn đến lợi ích đôi bên cùng có lợi. Do đó, đề tài này không chỉ có ý nghĩa học thuật mà còn mang lại giá trị thực tiễn, giúp doanh nghiệp tối ưu hóa chuỗi cung ứng và duy trì lợi thế cạnh tranh.

2.Phạm vi ứng dụng:

- **Phân tích chuỗi cung ứng:** Sử dụng dữ liệu để đánh giá hiệu suất và chất lượng của các nhà cung cấp, từ đó tối ưu hóa chuỗi cung ứng và lựa chọn các đối tác cung cấp đáng tin cậy.
- **Đánh giá chất lượng sản phẩm:** Xác định chất lượng sản phẩm từ các nhà cung cấp thông qua phân tích dữ liệu về tỷ lệ lỗi, độ ổn định của sản phẩm và các chỉ số khác liên quan đến chất lượng.
- **Dự đoán rủi ro:** Sử dụng dữ liệu quá khứ để dự đoán các rủi ro có thể phát sinh từ các nhà cung cấp kém chất lượng, giúp doanh nghiệp có các biện pháp phòng ngừa kịp thời.
- **Tối ưu hóa lựa chọn nhà cung cấp:** Áp dụng các phương pháp khai thác dữ liệu để phân loại và đánh giá các nhà cung cấp, từ đó đưa ra quyết định chọn lựa nhà cung cấp phù hợp nhất với yêu cầu về chất lượng, giá cả và thời gian giao hàng.
- **Quản lý hiệu suất nhà cung cấp:** Phân tích dữ liệu để theo dõi và đánh giá hiệu suất của các nhà cung cấp theo thời gian, giúp phát hiện các xu hướng và cải thiện mối quan hệ hợp tác.
- **Kết luận:** Việc áp dụng các phương pháp khai thác dữ liệu trong việc đánh giá chất lượng nhà cung cấp giúp doanh nghiệp không chỉ tối ưu hóa quá trình lựa chọn và quản lý nhà cung cấp mà còn giảm thiểu rủi ro và nâng cao hiệu quả hoạt động.

CHƯƠNG 2: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

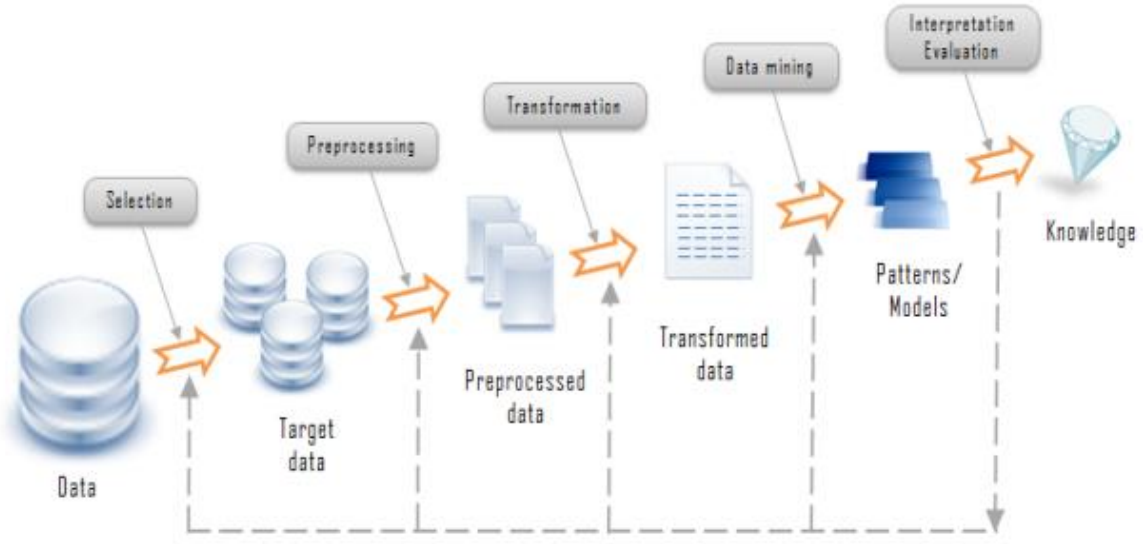
1.Khai phá dữ liệu (Data Mining)

- Khai phá dữ liệu (Data mining) là một khái niệm ra đời vào những năm cuối của thập kỷ 80. Nó bao hàm một loạt các kỹ thuật nhằm phát hiện ra các thông tin có giá trị tiềm ẩn trong các tập dữ liệu lớn trong thực tế. Về bản chất, khai phá dữ liệu liên quan đến việc phân tích các dữ liệu và sử dụng các kỹ thuật để tìm ra các mẫu hình có tính chính quy (regularities) từ các tập dữ liệu lớn nhằm mục đích dự đoán các xu thế, các hành vi trong tương lai, hoặc tìm kiếm những tập thông tin hữu ích mà bình thường không thể nhận diện được. Năm 1989, Fayyad, Piatetsky- Shapiro và Smyth đã dùng khái niệm Phát hiện tri thức trong cơ sở dữ liệu (Knowledge Discovery in Database-KDD) để chỉ toàn bộ quá trình phát hiện các tri thức có ích từ các tập dữ liệu lớn. Trong đó, khai phá dữ liệu là một bước đặc biệt trong toàn bộ quá trình, sử dụng các giải thuật đặc biệt để chiết xuất ra các mẫu (pattern) hay các mô hình từ dữ liệu.
- Khai phá dữ liệu nhấn mạnh hai khía cạnh chính đó là khả năng trích xuất thông tin có ích tự động (Automated) và bán tự động (Semi - Automated) mang tính dự đoán (Predictive). Khai phá dữ liệu là một lĩnh vực liên ngành, liên quan chặt chẽ đến các lĩnh vực sau:

- Statistics (Thống kê) : là một số đo cho một thuộc tính nào đó của một tập mẫu. Mỗi giá trị thống kê được tính bằng một hàm nào đó và thông tin của một thống kê mang tính đại diện cho thông tin của tập mẫu mang lại.
- Machine Learning (Máy học): là một phương pháp để tạo ra các chương trình máy tính bằng việc phân tích các tập dữ liệu. Máy học có liên quan lớn đến thống kê, vì cả hai lĩnh vực đều nghiên cứu việc phân tích dữ liệu, nhưng khác với thống kê, học máy tập trung vào sự phức tạp của các giải thuật trong việc thực thi tính toán
- Databases technology (Công nghệ CSDL): kho thông tin về một chủ đề, được tổ chức hợp lý để dễ dàng quản lý và truy tìm.
- Visualization (Sự trực quan): Biểu diễn giúp dữ liệu dễ hiểu, dễ sử dụng, thuận tiện cho việc tạo các báo cáo, tìm ra các tri thức phục vụ việc ra quyết định và dự đoán của nhà quản lý

2.Quy trình khai phá dữ liệu

Quy trình khai phá dữ liệu là một chuỗi lặp và tương tác gồm các bước (giai đoạn) bắt đầu với dữ liệu thô (raw data) và kết thúc với tri thức (Knowledge of interest) đáp ứng được sự quan tâm của người sử dụng



Hình 1. 1 Quy trình khai phá dữ liệu.

1. Data cleaning (làm sạch dữ liệu): loại bỏ nhiễu và phần tử dữ liệu không nhất quán.
2. Data integration (tích hợp dữ liệu): tích hợp dữ liệu từ nhiều nguồn khác nhau vào một kho dữ liệu.
3. Data selection (lựa chọn dữ liệu): trích chọn dữ liệu liên quan đến nhiệm vụ phân tích được lấy từ cơ sở dữ liệu.
4. Data transformation (biến đổi dữ liệu): Là quá trình chuyển đổi dữ liệu, nơi dữ liệu được hợp nhất thành các dạng thích hợp cho việc khai phá.
5. Data mining (khai phá dữ liệu): Là quá trình chính yếu nơi mà các kỹ thuật khai phá được sử dụng để phát hiện ra các mẫu (patterns).
6. Pattern evaluation (đánh giá mẫu): xác định các mô hình cần thiết đại diện cho tri thức (Knowledge) dựa trên các interestingness measures (tính dễ hiểu, phù hợp, hữu ích...).
7. Knowledge presentation (biểu diễn tri thức): Là quá trình sử dụng các công cụ trực quan hóa và các kỹ thuật biểu diễn tri thức để trình bày các tri thức được khai phá cho người sử dụng.

3.Quy trình xây dựng mô hình khai phá dữ liệu (Data Mining Model)

Việc thực hiện một DMM với đầy đủ 4 bước công việc chính của quá trình khai phá dữ liệu là:

Bước 1: Chuẩn bị dữ liệu (Data Preparation), trong bước này chúng ta thực hiện các công việc tiền xử lý dữ liệu theo yêu cầu của mô hình như trích chọn thuộc tính, rời rạc hóa dữ liệu và cuối cùng là chia dữ liệu nguồn (Data Source) thành 2 tập dữ liệu dùng để huấn luyện mô hình (Training Data) và kiểm tra mô hình (Testing data).

Bước 2: Xây dựng mô hình (Data Modeling), ta sử dụng Training Data vừa tạo ra để xây dựng mô hình.

Bước 3: Đánh giá mô hình (Validation), sau khi sử dụng Training Data để xây dựng mô hình, bây giờ ta sử dụng Testing Data để kiểm tra xem mô hình có đủ tốt để sử dụng hay không? (Nếu chưa đủ tốt thì phải sử dụng Training Data khác để huấn luyện lại). Có 3 kỹ thuật chính để kiểm tra mô hình đó là sử dụng Accuracy Chart (Lift Chart), Classification Matrix và Profit Chart.

Bước 4: Sử dụng mô hình để dự đoán dữ liệu trong tương lai (Model Usage), sau khi mô hình được kiểm tra (Testing) nếu độ chính xác đáp ứng yêu cầu thì có thể sử dụng model đã xây dựng vào dự đoán các dữ liệu chưa biết.

CHƯƠNG 3: GIỚI THIỆU KỸ THUẬT PHÂN CỤM TRONG KHAI PHÁ DỮ LIỆU

1.Chuẩn hoá dữ liệu Min-Max (Min-Max Scaling)

Chuẩn hóa Min-Max là một phương pháp để biến đổi dữ liệu sao cho tất cả các giá trị của các đặc trưng (features) nằm trong một khoảng nhất định, thường là từ 0 đến 1. Điều này giúp các đặc trưng có giá trị nằm trong cùng một khoảng, giúp các thuật toán học máy hoạt động hiệu quả hơn.

Phương pháp này sử dụng công thức sau:

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Trong đó:

- X là giá trị ban đầu của đặc trưng.
- Xmin:là giá trị nhỏ nhất của đặc trưng đó.
- Xmax: là giá trị lớn nhất của đặc trưng đó.
- Xscaled: là giá trị của đặc trưng sau khi chuẩn hóa.

Khi sử dụng MinMaxScaler với feature_range=(0, 1) trong thư viện sklearn, nó sẽ tự động áp dụng công thức trên cho các cột đã chọn để chuẩn hóa.

2.Giới thiệu bài toán phân cụm:

Phân cụm dữ liệu

- Phân cụm dữ liệu (Data Clustering) là một kỹ thuật quan trọng trong lĩnh vực khai phá dữ liệu và học máy, được sử dụng để chia tập dữ liệu thành các nhóm (cụm) sao cho các đối tượng trong cùng một nhóm có tính chất giống nhau nhiều hơn so với các đối tượng trong các nhóm khác. Đây là một phương pháp học không giám sát, tức là không cần nhãn trước (labels) cho dữ liệu.
- Phân cụm dữ liệu là quá trình phân chia một tập hợp các đối tượng dữ liệu thành các cụm (clusters) sao cho các đối tượng trong cùng một cụm có tính chất tương đồng với nhau nhiều hơn so với các đối tượng trong các cụm khác. Tính chất này thường được đo lường bằng một số hàm khoảng cách hoặc độ tương tự giữa các đối tượng.

Mục tiêu:

- **Khám phá cấu trúc dữ liệu:** Giúp phát hiện ra các mẫu, xu hướng, và mối quan hệ trong dữ liệu mà không cần thông tin trước về các nhãn.
- **Tóm tắt dữ liệu:** Giúp giảm độ phức tạp của dữ liệu bằng cách nhóm các đối tượng tương tự lại với nhau.
- **Tìm kiếm thông tin hữu ích:** Hỗ trợ việc ra quyết định dựa trên sự hiểu biết về cấu trúc của dữ liệu.
- **Phát hiện bất thường:** Nhận diện các đối tượng dữ liệu không thuộc về bất kỳ cụm nào, thường được coi là các điểm bất thường hoặc ngoại lệ.

Quá trình phân cụm dữ liệu

Quá trình phân cụm dữ liệu bao gồm một loạt các bước từ việc chuẩn bị dữ liệu cho đến việc đánh giá kết quả phân cụm. Dưới đây là các bước chi tiết trong quá trình phân cụm dữ liệu:

1. **Thu thập và chuẩn bị dữ liệu**:

    - **Thu thập dữ liệu**: Thu thập dữ liệu từ các nguồn khác nhau như cơ sở dữ liệu, tệp CSV, API, v.v.
    - **Tiền xử lý dữ liệu**: Làm sạch dữ liệu bằng cách loại bỏ các giá trị thiếu, xử lý dữ liệu bị nhiễu và chuẩn hóa dữ liệu để đảm bảo các biến có tỷ lệ so sánh được.

2. **Lựa chọn đặc trưng**:

    - **Lựa chọn và trích xuất đặc trưng**: Chọn ra các đặc trưng (features) quan trọng từ dữ liệu để phục vụ cho quá trình phân cụm. Có thể sử dụng kỹ thuật giảm chiều (dimensionality reduction) như PCA (Principal Component Analysis) để giảm số lượng đặc trưng nếu cần thiết.

3. **Lựa chọn thuật toán phân cụm**:

    - **Xem xét các thuật toán phân cụm**: Tùy thuộc vào dữ liệu và mục tiêu phân cụm, chọn thuật toán phân cụm phù hợp như K-means, Hierarchical clustering, DBSCAN, v.v.
    - **Thiết lập tham số**: Xác định các tham số cần thiết cho thuật toán (ví dụ: số lượng cụm k trong K-means).

4. **Thực hiện phân cụm**:

    - **Áp dụng thuật toán phân cụm**: Sử dụng thuật toán đã chọn để chia dữ liệu thành các cụm. Quá trình này có thể đòi hỏi tính toán lặp đi lặp lại cho đến khi đạt được kết quả phân cụm ổn định (convergence).

5. **Đánh giá kết quả phân cụm**:

    - **Đánh giá nội tại**: Sử dụng các chỉ số đánh giá như Sum of Squared Errors (SSE), Silhouette Score, Davies-Bouldin Index, v.v. để đánh giá chất lượng phân cụm dựa trên sự tương đồng và khác biệt giữa các cụm.
    - **Đánh giá ngoại tại**: Nếu có sẵn nhãn (labels) cho dữ liệu, có thể so sánh các cụm với các nhãn thực tế để đánh giá độ chính xác (accuracy), độ nhạy (recall), và độ chính xác (precision).

6. **Hiệu chỉnh và tối ưu hóa**:

    - **Điều chỉnh tham số**: Nếu kết quả chưa tốt, điều chỉnh tham số của thuật toán và thực hiện phân cụm lại.
    - **Thử các thuật toán khác**: Nếu cần, thử nghiệm với các thuật toán phân cụm khác để tìm ra thuật toán phù hợp nhất với dữ liệu và mục tiêu.

7. **Diễn giải và ứng dụng kết quả**:

    - **Diễn giải kết quả**: Phân tích và diễn giải các cụm đã tìm được, hiểu rõ ý nghĩa của mỗi cụm trong bối cảnh bài toán cụ thể.
    - **Ứng dụng kết quả**: Sử dụng các cụm để đưa ra các quyết định kinh doanh, cải thiện mô hình dự đoán, cá nhân hóa dịch vụ, v.v.

Quá trình phân cụm dữ liệu không chỉ đơn thuần là việc áp dụng thuật toán mà còn bao gồm các bước tiền xử lý, lựa chọn đặc trưng, đánh giá và tối ưu hóa để đảm bảo kết quả phân cụm đạt chất lượng cao và có ý nghĩa thực tiễn.

So sánh các mô hình phân cụm

- Trong từng ứng dụng cụ thể cần lựa chọn mô hình phân lớp phù hợp.
- Việc lựa chọn đó căn cứ vào sự so sánh các mô hình phân lớp với nhau, dựa trên các tiêu chuẩn sau:
  - Độ chính xác dự đoán (predictive accuracy) : độ chính xác là khả năng của mô hình để dự đoán chính xác nhãn lớp của dữ liệu mới hay dữ liệu chưa chết.
  - Tốc độ (speed) : tốc độ là những chi phí tính toán liên quan đến quá trình tạo ra và sử dụng mô hình
  - Sức mạnh (robustness) sức mạnh là khả năng mô hình tạo ra những dự đoán đúng từ những dữ liệu noise hay dữ liệu với những giá trị thiếu.
  - Tính hiểu được (interpretability) : tính hiểu được là mức độ hiểu và hiểu rõ những kết quả sinh ra bởi mô hình đọc.
  - Tính đơn giản (simplicity) : tính đơn giản liên quan đến kích thước của cây quyết định hay độ cô đọng của các luật. Trong các tiêu chuẩn trên, khả năng mở rộng của mô hình phân lớp được nhấn mạnh và chú trọng phát triển, đặc biệt với cây quyết định.

3.Thuật toán K-Means

3.1.Khái niệm:

Thuật toán K-Means là một thuật toán phân cụm (clustering) phổ biến trong học máy không giám sát (unsupervised learning). Mục tiêu của thuật toán là phân chia n điểm dữ liệu thành k cụm sao cho mỗi điểm dữ liệu thuộc về cụm có trung tâm gần nhất (centroid). Thuật toán này cố gắng tối thiểu hóa tổng bình phương khoảng cách từ mỗi điểm dữ liệu đến trung tâm của cụm mà nó thuộc về.

3.2.Quy trình hoạt động của K-Means

- **Khởi Tạo:**
  - Chọn k điểm ngẫu nhiên từ dữ liệu để làm trung tâm ban đầu của các cụm (centroids).
- **Phân Cụm:**
  - Gán mỗi điểm dữ liệu vào cụm có trung tâm gần nhất. Điều này được thực hiện bằng cách tính khoảng cách Euclid giữa điểm dữ liệu và mỗi trung tâm cụm và gán điểm dữ liệu vào cụm có khoảng cách ngắn nhất.
- **Cập Nhật Trung Tâm Cụm:**
  - Sau khi tất cả các điểm dữ liệu đã được gán vào cụm, tính toán lại trung tâm của mỗi cụm. Trung tâm mới của mỗi cụm là trung bình cộng của tất cả các điểm dữ liệu trong cụm đó.
- **Lặp Lại:**
  - Lặp lại các bước 2 và 3 cho đến khi trung tâm các cụm không thay đổi (hoặc thay đổi rất ít), hoặc đạt đến số lần lặp tối đa. Điều này có nghĩa là thuật toán đã hội tụ.

3.3.Ưu điểm và nhược điểm K-Means:

- **Ưu điểm:**
  - Đơn giản và dễ hiểu.
  - Tính toán nhanh, hiệu quả với các tập dữ liệu lớn.
- **Nhược điểm:**
  - Phải xác định số cụm kkk trước.
  - Không đảm bảo tìm được lời giải tối ưu toàn cục.
  - Nhạy cảm với giá trị khởi tạo và nhiễu trong dữ liệu.

CHƯƠNG 4 : TÌM HIỂU VỀ CÔNG CỤ

1.Weka:

1.1.Khái niệm:

- Giới thiệu Khai phá dữ liệu (Data Mining) và học máy (Machine Learning) là những lĩnh vực khá khó để khám phá và nghiên cứu. Do đó, nhiều phần mềm đã ra đời với mục tiêu là giúp cho người dùng có thể dễ dàng nghiên cứu các bài toán trong những lĩnh vực khó nhằn này. Những phần mềm đó có thể kể đến như Matlab, Orange, KNIME hay RapidMiner. Trong bài viết này, chúng em sẽ đề cập đến phần mềm WEKA, một phần mềm mã nguồn mở tuyệt vời dành cho khai phá dữ liệu.
- WEKA - Waikato Environment for Knowledge Analysis, là bộ phần mềm học máy, mã nguồn mở, do đại học Waikato phát triển bằng JAVA, nhằm phục vụ cho các nhiệm vụ chuyên về khai phá dữ liệu.
- WEKA chứa các công cụ phục vụ cho tiền xử lý dữ liệu, phân loại, hồi quy, phân cụm, các luật liên quan và trực quan hóa. Nó cũng phù hợp cho việc phát triển, xây dựng các mô hình học máy và có khả năng chạy được trên nhiều hệ điều hành khác nhau như Windows, Mac, Linux.

1.2.Lịch sử phát triển:

- 1993: Đại học Waikato của New Zealand bắt đầu xây dựng phiên bản đầu tiên của phần mềm Weka.
- 1997: Xây dựng lại Weka từ đầu bằng ngôn ngữ Java, có cài đặt các thuật toán mô hình hóa.
- 2005: Weka xuất sắc nhận được giải thưởng SIGKDD Data Mining and Knowledge Discovery Service Award.
- 2007: Phần mềm đứng thứ 241 trong những phần mềm được tải nhiều nhất trên Sourceforge.net

1.3.Các chức năng chính:

- Những tính năng vượt trội trong Weka có thể kể đến là:
  - Mã nguồn mở
  - Hỗ trợ các thuật toán học máy (machine learning) và khai phá dữ liệu
  - Trực quan hóa, dễ dàng xây dựng các ứng dụng thực nghiệm
  - Do sử dụng JVM nên Weka độc lập với môi trường Kiến trúc trong thư viện Weka bao gồm hơn 600 class và được tổ chức thành 10 package. Chính vì thế, người sử dụng có thể dùng trực tiếp trên phần mềm hoặc sử dụng những class này làm bộ thư viện để phát triển các ứng dụng của riêng mình.

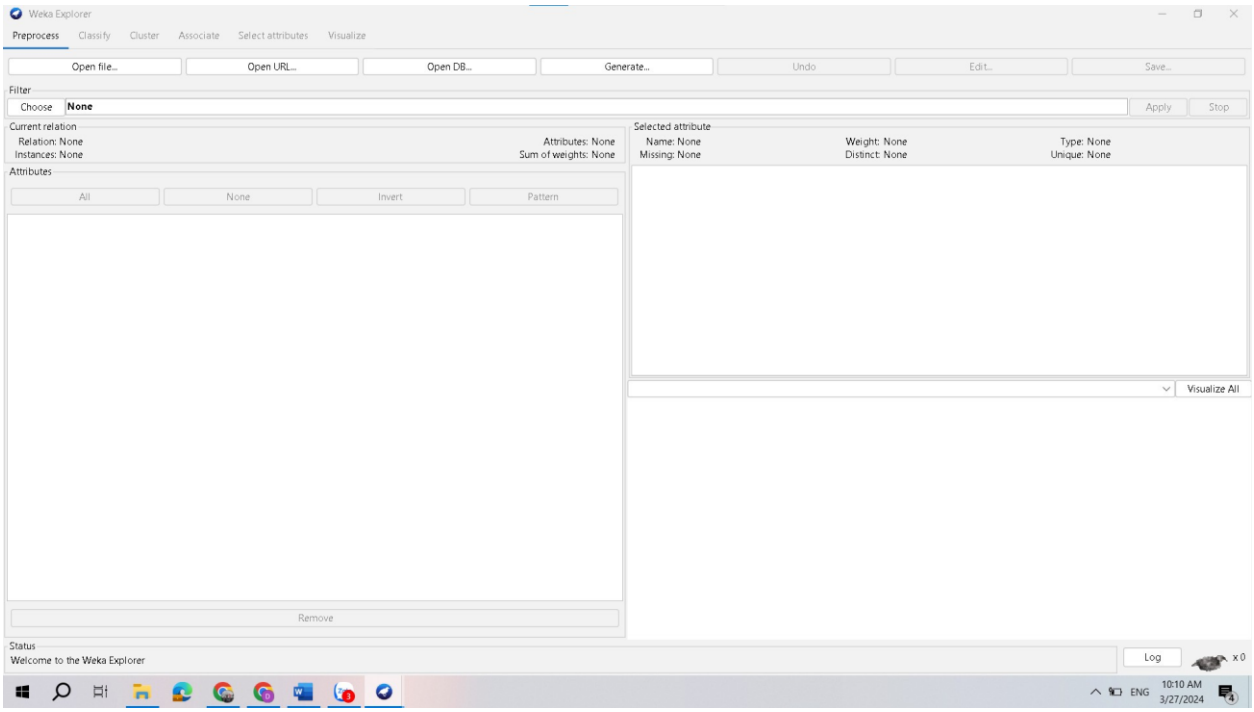
1.4.Giao diện và cách sử dụng:

Các môi trường chính:

Hình 4. 1 Giao diện khởi động WEKA

- Simple CLI Giao diện đơn giản kiểu dòng lệnh (như MS-DOS)
- **Explore** là môi trường cho phép chúng ta sử dụng tất cả các khả năng của WEKA để khai phá dữ liệu, cho phép thực nghiệm các nhiệm vụ khai thác dữ liệu thường gặp.

➢

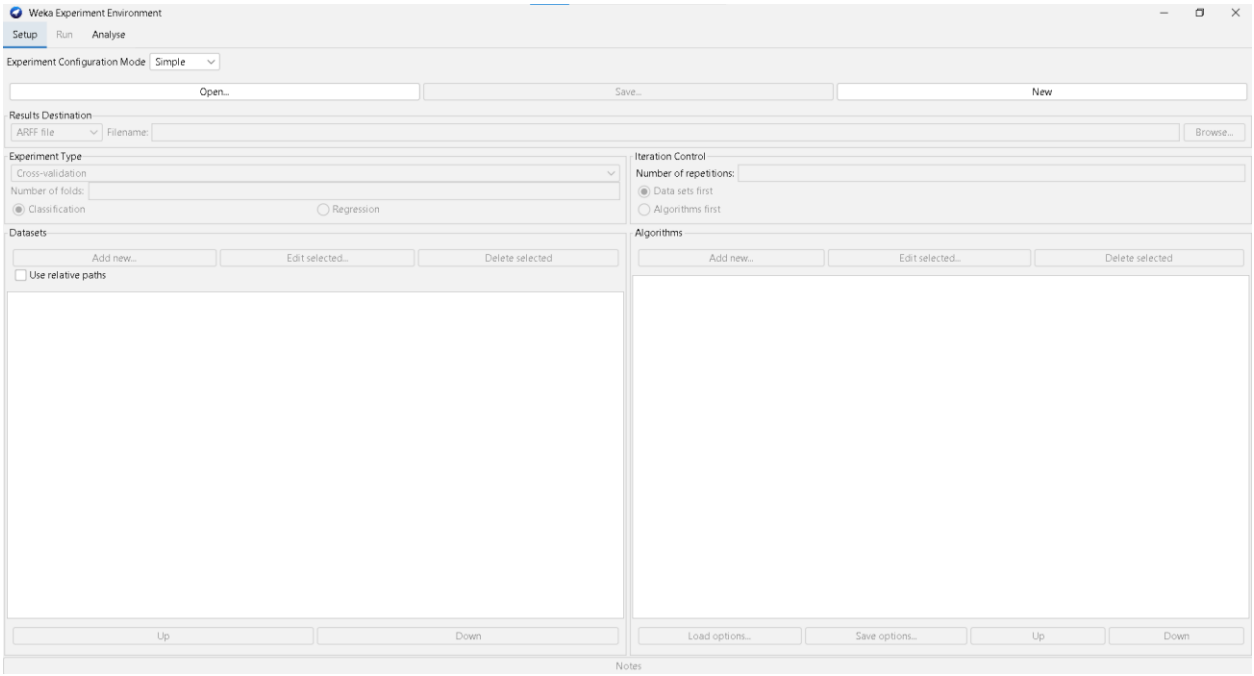


Hình 4. 2 Gio diện môi trường Explorer

Mô tả các chức năng:

- Preprocess để chọn và thay đổi (xử lý) dữ liệu làm việc
- Classify để huấn luyện các mô hình máy học(phân loại hoặc hồi quy/ dự đoán)
- Cluster để học các nhóm từ dữ liệu (phân cụm)
- Associate để khám phá các luật kết hợp từ dữ liệu
- Select attributes để xác định và lựa chọn các thuộc tính có liên quan (quan trọng) nhất của dữ liệu
- Visualize để xem và hiển thị biểu đồ tương tác 2 chiều đối với dữ liệu

➢ **Experimenter** môi trường cho phép thực hiện các thí nghiệm và thực hiện các kiểm tra thống kê (statistical tests) giữa các mô hình máy học



Hình 4. 3 Giao diện môi trường Experimenter

➢ **KnowledgeFlow** môi trường cho phép bạn tương tác đồ họa theo kiểu kéo/thả để thiết kế các bước (các thành phần) của một thí nghiệm.

➢ Các bộ phân lớp trong WEKA-Explorer:

- Các bộ phân lớp (Classifiers) của WEKA tương ứng với các mô hình dự đoán các đại lượng kiểu định danh (phân lớp) hoặc các đại lượng kiểu số (hồi quy/dự đoán).
- Các kỹ thuật phân lớp được hỗ trợ bởi WEKA:
  - Naive Bayes classifier and Bayesian networks.
  - Decision trees.
  - Instance-based classifiers.
  - Support vector machines.
  - Neural networks
- Lựa chọn một bộ phân lớp (classifier).
- Lựa chọn các tùy chọn cho việc kiểm tra (test options):
  - Use training set. Bộ phân loại học được sẽ được đánh giá trên tập học •
  - Supplied test set. Sử dụng một tập dữ liệu khác (với tập học) để cho việc đánh giá
  - Cross-validation. Tập dữ liệu sẽ được chia đều thành k tập (folds) có kích thước xấp xỉ nhau, và bộ phân loại học được sẽ được đánh giá bởi phương pháp cross-validation
  - Percentage split. Chỉ định tỷ lệ phân chia tập dữ liệu đối với việc đánh giá - More options…
  - Output model Output model. Hiển thị bộ phân lớp học được
  - Output per-class stats. Hiển thị các thông tin thống kê về precision/recall đối với mỗi lớp
  - Output entropy evaluation measures Output entropy evaluation measures. Hiển thị đánh giá độ hỗn tạp (entropy) của tập dữ liệu

- Output confusion matrix. Hiển thị thông tin về ma trận lỗi phân lớp ( ) confusion matrix) đối với phân lớp học được
- Store predictions for visualization. Các dự đoán của bộ phân lớp được lưu lại trong bộ nhớ, để có thể được hiển thị sau đó
- Output predictions. Hiển thị chi tiết các dự đoán đối với tập kiểm tra
- Cost-sensitive evaluation. Các lỗi (của bộ phân lớp) được xác định dựa trên ma trận chi phí (cost matrix) chỉ định
- Random seed for XVal / % Split : Chỉ định giá trị random seed được sử dụng cho quá trình lựa chọn ngẫu nhiên các ví dụ cho tập kiểm tra - Classifier output hiển thị các thông tin quan trọng
- Run information : Các tùy chọn đối với mô hình học tên, tên của tập dữ liệu, số lượng các ví dụ, các thuộc tính, và f.f. thí nghiệm
- Classifier model (full training set) : Biểu diễn (dạng text) của bộ phân lớp học được.
- Predictions on test data : Thông tin chi tiết về các dự đoán của bộ phân lớp đối với tập kiểm tra
- Summary : Các thống kê về mức độ chính xác của bộ phận lớp, đối với f.f. thí nghiệm đã chọn.
- Detailed Accuracy By Class : Thông tin chi tiết về mức độ chính xác của bộ phận lớp đối với mỗi lớp .
- Confusion Matrix : Các thành phần của ma trận này thể hiện số lượng các ví dụ kiểm tra (test instances) được phân lớp đúng và bị phân lớp sai. - Result list cung cấp một số chức năng hữu ích:
- Save model Save model : Lưu lại mô hình tương ứng với bộ phân lớp học được vào trong một tập tin nhị phân (binary file)
- Load model : Đọc lại một mô hình đã được học trước đó từ một tập tin nhị phân
- Re-evaluate model on current test set : Đánh giá một mô hình (bộ phân lớp) học được trước đó đối với tập kiểm tra (test set) hiện tại
- Visualize classifier errors : Hiển thị cửa sổ biểu đồ thể hiện các kết quả của việc phân lớp. Các ví dụ được phân lớp chính xác sẽ được biểu diễn bằng ký hiệu bởi dấu chéo (x), còn các ví dụ bị phân lớp sau sẽ được biểu diễn bằng ký hiệu ô vuông ()

Các chức năng chính:

1.Tiền xử lí dữ liệu

2.Huấn luyện các mô hình máy học(phân lớp)

3.Gom nhóm

2.Python:

2.1.Khái niệm:

Python là một ngôn ngữ lập trình được sử dụng rộng rãi trong các ứng dụng web, phát triển phần mềm, khoa học dữ liệu và máy học (ML). Các nhà phát triển sử dụng Python vì nó hiệu quả, dễ học và có thể chạy trên nhiều nền tảng khác nhau. Phần mềm Python được tải xuống miễn phí, tích hợp tốt với tất cả các loại hệ thống và tăng tốc độ phát triển.

2.2.Lợi ích

Những lợi ích của Python bao gồm:

- Các nhà phát triển có thể dễ dàng đọc và hiểu một chương trình Python vì ngôn ngữ này có cú pháp cơ bản giống tiếng Anh.
- Python giúp cải thiện năng suất làm việc của các nhà phát triển vì so với những ngôn ngữ khác, họ có thể sử dụng ít dòng mã hơn để viết một chương trình Python.
- Python có một thư viện tiêu chuẩn lớn, chứa nhiều dòng mã có thể tái sử dụng cho hầu hết mọi tác vụ. Nhờ đó, các nhà phát triển sẽ không cần phải viết mã từ đầu.
- Các nhà phát triển có thể dễ dàng sử dụng Python với các ngôn ngữ lập trình phổ biến khác như Java, C và C++.
- Cộng đồng Python tích cực hoạt động bao gồm hàng triệu nhà phát triển nhiệt tình hỗ trợ trên toàn thế giới. Nếu gặp phải vấn đề, bạn sẽ có thể nhận được sự hỗ trợ nhanh chóng từ cộng đồng.
- Trên Internet có rất nhiều tài nguyên hữu ích nếu bạn muốn học Python. Ví dụ: bạn có thể dễ dàng tìm thấy video, chỉ dẫn, tài liệu và hướng dẫn dành cho nhà phát triển.
- Python có thể được sử dụng trên nhiều hệ điều hành máy tính khác nhau, chẳng hạn như Windows, macOS, Linux và Unix.

CHƯƠNG 5: TRIỂN KHAI ỨNG DỤNG GOM CỤM DỮ LIỆU & ĐÁNH GIÁ KHO DỮ LIỆU

1.Thu thập dữ liệu

1.1.Mục đích:

Mục đích của việc thu thập dữ liệu để xây dựng và quản lý kho dữ liệu bao gồm nhiều khía cạnh quan trọng giúp tổ chức tối ưu hóa hoạt động kinh doanh và ra quyết định chiến lược. Cụ thể, mục đích của việc này có thể bao gồm:

- **Tích Hợp Dữ Liệu:** Tập hợp dữ liệu từ nhiều nguồn khác nhau vào một kho dữ liệu tập trung, giúp đảm bảo tính nhất quán và toàn vẹn của dữ liệu.
- **Cải Thiện Chất Lượng Dữ Liệu:** Làm sạch và chuẩn hóa dữ liệu để loại bỏ sai sót và đảm bảo rằng dữ liệu trong kho dữ liệu là chính xác và đáng tin cậy.
- **Hỗ Trợ Quyết Định:** Cung cấp một cơ sở dữ liệu chung và đáng tin cậy để hỗ trợ các nhà quản lý và lãnh đạo đưa ra quyết định dựa trên dữ liệu.
- **Phân Tích Đa Chiều:** Cung cấp khả năng phân tích dữ liệu từ nhiều góc độ khác nhau (OLAP), cho phép người dùng tạo các báo cáo và biểu đồ phức tạp để hiểu rõ hơn về dữ liệu.
- **Lưu Trữ Lâu Dài:** Lưu trữ dữ liệu lịch sử trong một khoảng thời gian dài để phục vụ cho việc phân tích xu hướng và dự báo trong tương lai.
- **Tiết Kiệm Thời Gian và Chi Phí:** Giảm thời gian và chi phí liên quan đến việc truy xuất và phân tích dữ liệu từ nhiều nguồn khác nhau bằng cách tập trung dữ liệu vào một kho duy nhất.
- **Hỗ Trợ Báo Cáo và BI (Business Intelligence):** Tạo ra các báo cáo, dashboard và công cụ phân tích BI để cung cấp thông tin chi tiết và hỗ trợ việc ra quyết định kinh doanh.
- **Cải Thiện Hiệu Suất Truy Vấn:** Tối ưu hóa việc truy vấn và truy xuất dữ liệu so với việc lấy dữ liệu từ các hệ thống giao dịch trực tiếp, nhờ đó cải thiện tốc độ và hiệu suất phân tích.
- **Phân Tích Rủi Ro:** Đánh giá các rủi ro tiềm ẩn và lập kế hoạch giảm thiểu rủi ro dựa trên dữ liệu thu thập được.
- **Đảm Bảo Tuân Thủ và Báo Cáo:** Thu thập dữ liệu để đảm bảo tuân thủ các quy định pháp luật và tiêu chuẩn ngành, cũng như để báo cáo cho các bên liên quan.

Tóm lại, việc thu thập dữ liệu để xây dựng và quản lý kho dữ liệu giúp tổ chức có một nền tảng vững chắc cho các hoạt động phân tích, báo cáo, và ra quyết định, từ đó nâng cao hiệu quả kinh doanh và khả năng cạnh tranh trên thị trường.

1.2.Đặc tả kho dữ liệu:

Mục đích:

Đặc tả kho dữ liệu (data warehouse specification) là một tài liệu chi tiết mô tả các yêu cầu, cấu trúc, và quy trình liên quan đến việc xây dựng và quản lý kho dữ liệu. Đặc tả này bao gồm nhiều khía cạnh kỹ thuật và nghiệp vụ để đảm bảo rằng kho dữ liệu đáp ứng được mục tiêu của tổ chức.

Kho dữ liệu bao gồm các thuộc tính:

- Nhà cung cấp:

  Nhà cung cấp (Nha Cung Cap) là một cá nhân, công ty hoặc tổ chức cung cấp hàng hóa hoặc dịch vụ cho một doanh nghiệp hoặc tổ chức khác. Nhà cung cấp đóng vai trò quan trọng trong chuỗi cung ứng, cung cấp các yếu tố đầu vào cần thiết để sản xuất sản phẩm hoặc cung cấp dịch vụ cuối cùng.

- Sản Phẩm:

  Sản phẩm là một đối tượng hoặc dịch vụ được tạo ra nhằm đáp ứng nhu cầu và mong muốn của người tiêu dùng hoặc thị trường. Nó có thể ở dạng hữu hình (như hàng hóa vật chất) hoặc vô hình (như dịch vụ). Sản phẩm là kết quả của quá trình sản xuất hoặc cung cấp dịch vụ và được bán hoặc cung cấp cho khách hàng

  Mỗi sản phẩm có tên gọi riêng biệt dùng để nhận diện và phân biệt sản phẩm đó với các sản phẩm khác trên thị trường. Tên sản phẩm thường phản ánh đặc điểm, chức năng, hoặc thương hiệu của sản phẩm và giúp khách hàng dễ dàng nhận biết và ghi nhớ.

- Mã Đơn Hàng:

  Mã đơn hàng là một chuỗi ký tự hoặc số duy nhất được gán cho mỗi đơn hàng khi nó được tạo ra. Mã này giúp theo dõi, quản lý và phân biệt từng đơn hàng riêng lẻ trong hệ thống quản lý đơn hàng của doanh nghiệp

- Số Lượng Đặt:

  Số lượng đặt hàng là số lượng sản phẩm hoặc dịch vụ mà khách hàng yêu cầu mua khi tạo một đơn hàng. Nó phản ánh số lượng đơn vị sản phẩm mà khách hàng muốn nhận hoặc thanh toán trong giao dịch

- Số Lượng Lỗi:

  Số lượng lỗi là số lượng sản phẩm hoặc hàng hóa bị lỗi trên mỗi đơn hàng

- Tỉ Lệ Lỗi:

  Tỉ lệ sản phẩm bị lỗi trên tổng số sản phẩm đơn hàng

- Chiết Khấu:

  Chiết khấu đơn hàng là khoản giảm giá áp dụng trực tiếp vào giá trị của một đơn hàng cụ thể. Đây là một hình thức ưu đãi mà người bán cung cấp cho khách hàng, thường dựa trên các yếu tố như số lượng mua, giá trị đơn hàng, thời gian thanh toán, hoặc các chương trình khuyến mãi đặc biệt.

- Quy Mô Nhà Cung Cấp:

  Quy mô nhà cung cấp là kích thước, năng lực và phạm vi hoạt động của nhà cung cấp trong việc cung cấp hàng hóa hoặc dịch vụ, bao gồm các yếu tố như sản lượng, tài chính, nhân sự và khả năng phân phối.

- Thời Gian Giao Hàng:

  Thời gian giao hàng là khoảng thời gian từ khi khách hàng đặt hàng cho đến khi sản phẩm hoặc dịch vụ được giao đến tay khách hàng. Thời gian này có thể thay đổi tùy thuộc vào các yếu tố như phương thức vận chuyển, khoảng cách, tính khả dụng của sản phẩm, và các chính sách của nhà cung cấp.

- Vị Trí:

  Vị trí của nhà cung cấp là địa điểm hoặc khu vực mà nhà cung cấp hoạt động hoặc đặt trụ sở, có thể là quốc gia, thành phố, hoặc vùng lãnh thổ cụ thể. Vị trí này ảnh hưởng đến khả năng cung cấp hàng hóa, chi phí vận chuyển, thời gian giao hàng và khả năng đáp ứng nhu cầu của khách hàng.

- Dịch Vụ Khách Hàng:

  Dịch vụ khách hàng là các hoạt động hỗ trợ và chăm sóc khách hàng trước, trong và sau khi mua sản phẩm hoặc dịch vụ, nhằm đảm bảo sự hài lòng và giải quyết các vấn đề của khách hàng.

Mô tả kho dữ liệu:

Thuộc Tính	Mô tả	Giá trị
Nha Cung Cap	Tên Nhà Cung Cấp sản phẩm	B_MOBI,C_MOBI,H_MOBI,SO_MOBI ET_MOBI
San Pham	Tên sản phẩm trong đơn hàng	O cung,Loa,Ram,Ban Phim,Chuot
Ma Don Hang	Mã đơn hàng	
So Luong Dat	Số lượng đặt hàng trong đơn hàng	
So Luong Loi	Số lượng lỗi là số sản phẩm bị lỗi trong đơn hàng	
Ti Le Loi	Tỉ lệ sản phẩm bị lỗi trên tổng số sản phẩm đơn hàng	SoLuongLoi / SoLuongDat
Chiet Khau	Chiết khấu trên đơn hàng	Chiết khấu từ 0-10%
Quy Mo	Quy mô Nhà Cung Cấp	1: Quy mô Nhỏ 2:Quy mô Trung Bình 3:Quy mô Lớn
Thoi Gian Giao Hang	Thời gian giao hàng kể từ ngày đặt	1:Nhanh 2:Trung Bình 3:Chậm
Vi Tri	Vị trí địa lí của Nhà Cung Cấp	1:Vị trí Xa 2:Vị trí Trung Bình 3:Vị trí Gần
Dich Vu Khach Hang	Dịch vụ hỗ trợ trong quá trình đặt hàng,và sau khi nhận hàng	1: Dịch vụ khách hàng Tôt 2: Dịch vụ khách hàng Trung Bình 3: Dịch vụ khách hàng Kem

1.3.Chuẩn bị dữ liệu cho mô hình

File dữ liệu gốc được lưu trong file : dulieugoc.csv

A	B	C	D	E	F	G	H	I	J
Nha Cung Cap	San Pham	So Luong Dat	So Luong Loi	Ti Le Loi	Chiet Khau	Thoi Gian Giao Hang	Quy Mô	Vi Tri	Dich Vu Khach Hang
B_MOBI	O cung	41	0	0	9	1	3	3	1
B_MOBI	Loa	80	0	0	8	1	1	1	1
B_MOBI	Loa	98	0	0	8	1	3	3	1
B_MOBI	O cung	98	0	0	9	1	3	3	1
B_MOBI	Loa	97	0	0	9	1	2	2	1
B_MOBI	Loa	15	0	0	8	1	3	3	1
B_MOBI	Loa	88	1	0.011	9	1	3	3	1
B_MOBI	Loa	80	1	0.013	10	1	3	3	1
B_MOBI	Loa	77	1	0.013	7	1	3	3	1
B_MOBI	Ban Phim	60	1	0.017	7	1	1	1	1
B_MOBI	Ram	56	1	0.018	8	1	2	2	1
B_MOBI	Ram	89	2	0.022	7	1	1	1	1
B_MOBI	Ban Phim	86	2	0.023	8	1	1	1	1
B_MOBI	Loa	95	3	0.032	10	1	1	1	1
B_MOBI	Loa	63	2	0.032	7	1	3	3	1
B_MOBI	Ban Phim	83	3	0.036	9	1	3	3	1
B_MOBI	Loa	78	3	0.038	7	1	1	1	1
B_MOBI	Ram	75	3	0.04	10	1	1	1	1
B_MOBI	Ban Phim	46	2	0.043	8	1	3	3	1
B_MOBI	Loa	68	3	0.044	10	1	3	3	1
B_MOBI	Loa	84	4	0.048	9	1	3	3	1
B_MOBI	Ban Phim	84	4	0.048	8	1	1	1	1
B_MOBI	Loa	62	3	0.048	10	1	3	3	1
B_MOBI	O cung	93	5	0.054	8	1	1	1	1
B_MOBI	Loa	74	5	0.068	10	1	2	2	1

Hình 5. 1 Bộ dữ liệu chưa chuẩn hoá

1.4.Chuẩn hoá dữ liệu:

Code Python chuẩn hoá dữ liệu

```
import pandas as pd
from sklearn.preprocessing import MinMaxScaler

# Đọc dữ liệu từ tệp
file_path = 'dulieugoc.csv'
data = pd.read_csv(file_path)

# Danh sách các cột cần chuẩn hóa
columns_to_normalize = [
    ‘Ti Le Loi’, 'Chiet Khau','Thoi Gian Giao Hang','Chat Luong San Pham', 'Vi Tri', 'Quy Mo','Dich Vu Khach Hang']

# Chuẩn hóa dữ liệu (Min-Max Normalization) sử dụng sklearn
scaler = MinMaxScaler(feature_range=(0, 1))
data_normalized = data.copy()
data_normalized[columns_to_normalize] = scaler.fit_transform(data[columns_to_normalize])

# Lưu dữ liệu chuẩn hóa vào file CSV
data_normalized.to_csv('dulieu_chuanhoa.csv', index=False)
print("Dữ liệu chuẩn hóa đã được lưu vào file 'dulieu_chuanhoa.csv'.")
```

Danh sách các cột được chuẩn hoá trong bộ dữ liệu gồm: 'So Luong Dat','So Luong Nhan', 'Chiet Khau','Thoi Gian Giao Hang','Chat Luong San Pham', 'Vi Tri', 'Quy Mo','Dich Vu Khach Hang'

Dữ liệu đã chuẩn hoá bằng MinMaxScaler(feature_range=(0, 1))

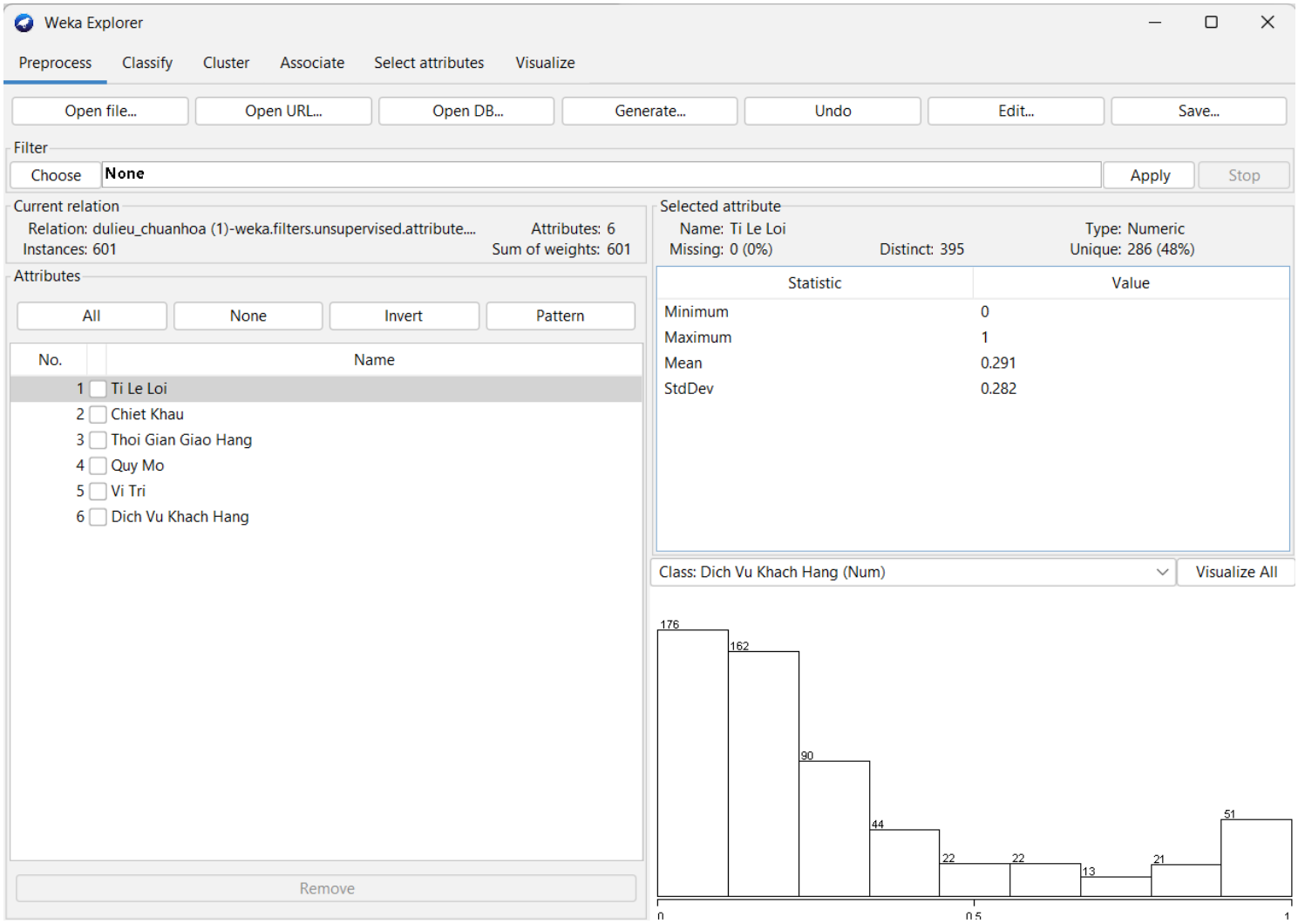
Nha Cung Cap	San Pham	So Luong Dat	So Luong Loi	Ti Le Loi	Chiet Khau	Thoi Gian Giao Hang	Quy Mo	Vi Tri	Dich Vu Khach Hang
B_MOBI	O cung	41	0	0	1	0	0.5	0.5	0
B_MOBI	Loa	80	0	0	1	0	1	1	0
B_MOBI	Loa	98	0	0	0.9	0	1	1	0
B_MOBI	O cung	98	0	0	0.9	0	1	1	0
B_MOBI	Loa	97	0	0	1	0	1	1	0
B_MOBI	Loa	15	0	0	0.8	0	0	0	0
B_MOBI	Loa	88	1	0.011931818	0.9	0	1	1	0
B_MOBI	Loa	80	1	0.013125	0.8	0	1	1	0
B_MOBI	Loa	77	1	0.013636364	1	0	1	1	0
B_MOBI	Ban Phim	60	1	0.0175	1	0	0.5	0.5	0
B_MOBI	Ram	56	1	0.01875	0.9	0	0.5	0.5	0
B_MOBI	Ram	89	2	0.023595506	0.9	0	1	1	0
B_MOBI	Ban Phim	86	2	0.024418605	0.8	0	1	1	0
B_MOBI	Loa	95	3	0.033157894	0.9	0	1	1	0
B_MOBI	Loa	63	2	0.033333334	0.9	0	0.5	0.5	0
B_MOBI	Ban Phim	83	3	0.037951807	0.8	0	1	1	0
B_MOBI	Loa	78	3	0.040384615	0.9	0	1	1	0
B_MOBI	Ram	75	3	0.042	0.9	0	1	1	0
B_MOBI	Ban Phim	46	2	0.045652174	1	0	0.5	0.5	0
B_MOBI	Loa	68	3	0.046323529	0.8	0	0.5	0.5	0
B_MOBI	Loa	84	4	0.05	0.8	0	1	1	0
B_MOBI	Ban Phim	84	4	0.05	0.9	0	1	1	0
B_MOBI	Loa	62	3	0.050806452	0.9	0	0.5	0.5	0
B_MOBI	O cung	93	5	0.056451613	0.8	0	1	1	0
B_MOBI	Loa	74	5	0.070945946	0.8	0	1	1	0

Hình 5. 2 Bộ dữ liệu sau khi chuẩn hoá

2.Tiến hành phân cụm dữ liệu:

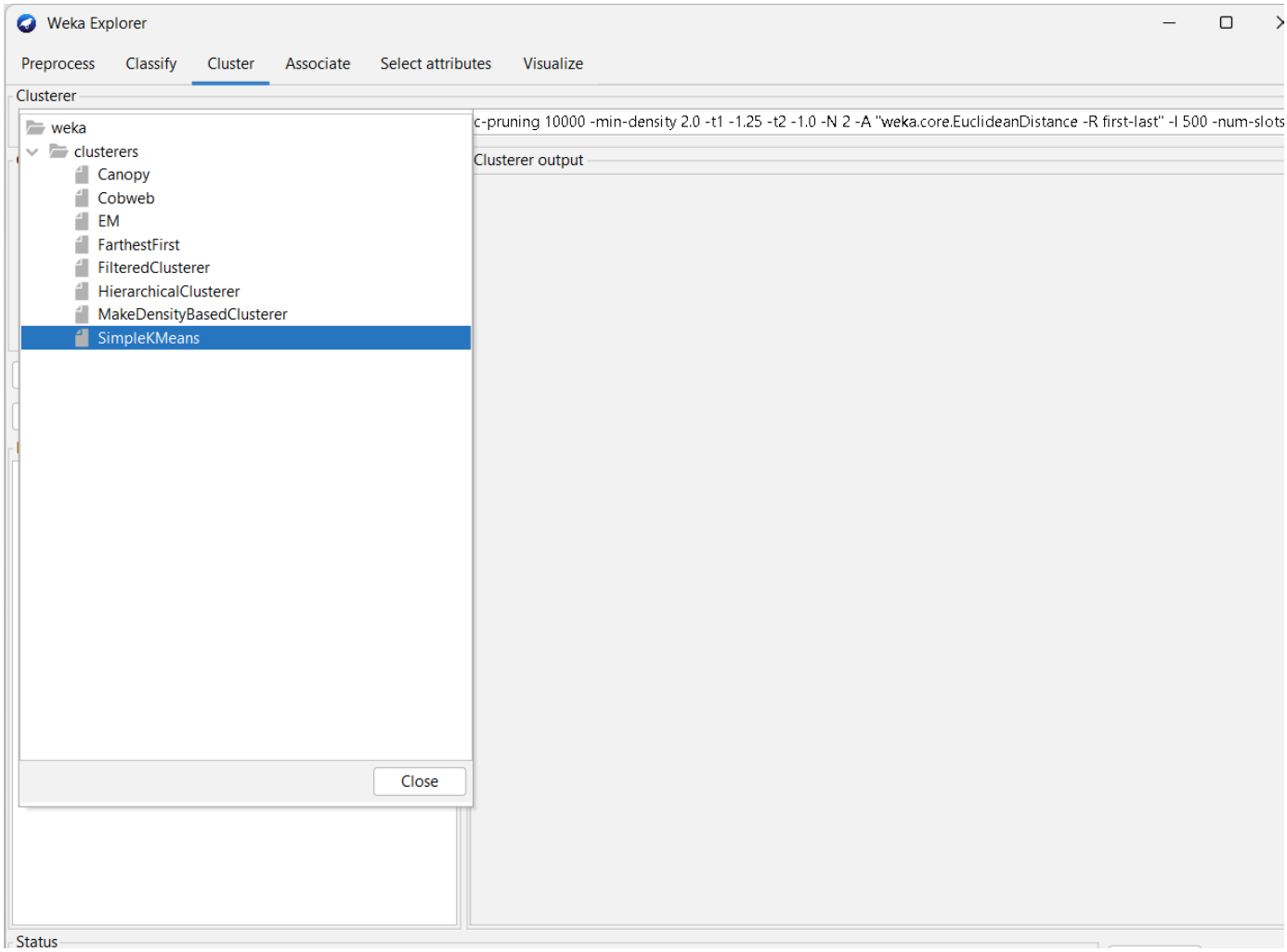
Khởi động WEKA, sau đó ta chọn button EXPLORER. Ở tab PREPROCESS, ta click vào nút Open files. Sau đó, tìm đến nơi lưu file để nạp dữ liệu vào

Sau khi nạp dữ liệu vào WEKA, ta tiến hành quan sát tập dữ liệu dựa vào Giao diện như sau :



Hình 5. 3 Hiển thị dữ liệu trên Weka Bộ dữ liệu

- Tại mục Attributes(danh sách các thuộc tính) có thứ tự dựa vào khai báo mà chúng ta nạp vào. Khi ta click vào từng thuộc tính thì tại mục Selected Attribute và phần biểu đồ trực quan cũng thay đổi tương ứng
- Chọn Cluster, sau đó chọn giải thuật mình muốn (SimpleKmeans)

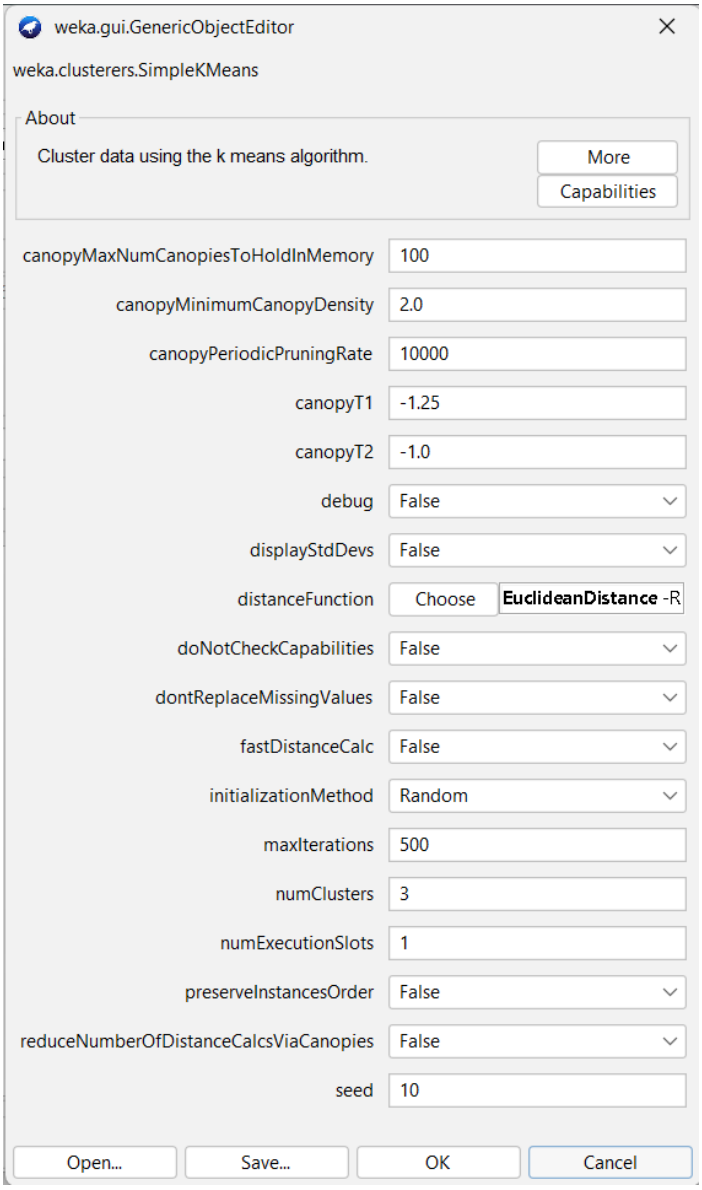


Hình 5. 4 Màn hình thiết lập giải thuật SimpleKmeans trên Weka

2.1.Tiến hành phân cụm trên bộ dữ liệu chuẩn:

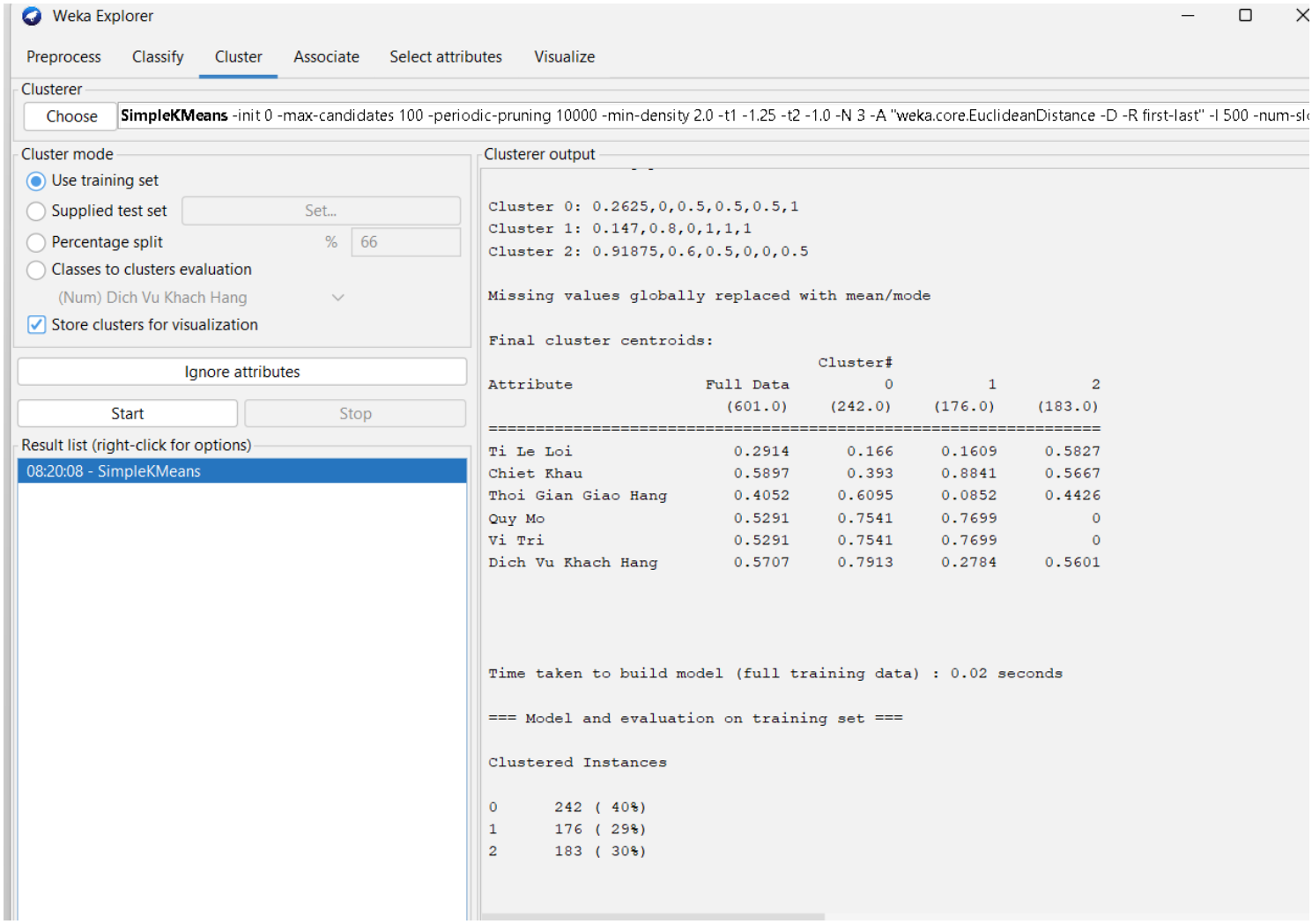
2.1.1.Cấu hình 1:

- Số cụm (k)=3
  - Bộ dữ liệu sẽ được chia làm 3 cụm.Mỗi cụm sẽ có vecto đặc trưng khách nhau nhằm gợi ý các nhà cung cấp có chất lượng khác nhau
- Số vòng lặp tối đa:500
  - Số lần lặp tối đa của thuật toán là 500 vòng.Nếu thuật toán hội tội trước khi đạt đến số vòng lặp này,quá trình sẽ kết thúc sớm
- Phương pháp khởi tạo:Random
  - Tâm cụm sẽ được chọn 1 cách ngẫu nhiên.Việc khởi tạo ngẫu nhiên này có thể làm thay đổi kết quả khi thay đổi giá trị seed
- Seed:10
  - Giá trị khởi tạo cho thuật toán.Việc chọn seed =10 đảm bảo tính tái lập,nghĩa là khi chạy thuật toán cũng cấp hình,kết quả sẽ không đổi.



Hình 5. 5.Cấu hình 1 trong phân cụm bộ dữ liệu chuẩn

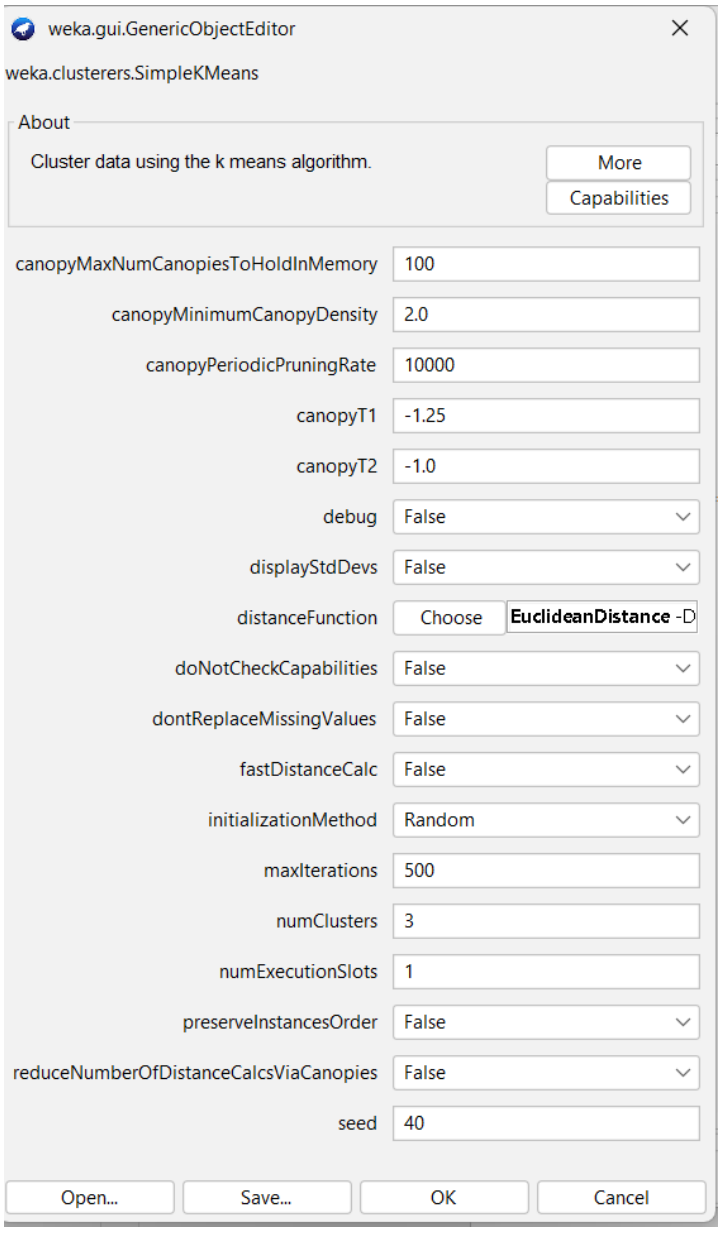
Kết quả phân cụm cấu hình 1 :



Hình 5. 6 Kết quả phân cụm bộ dữ liệu chuẩn cấu hình 1

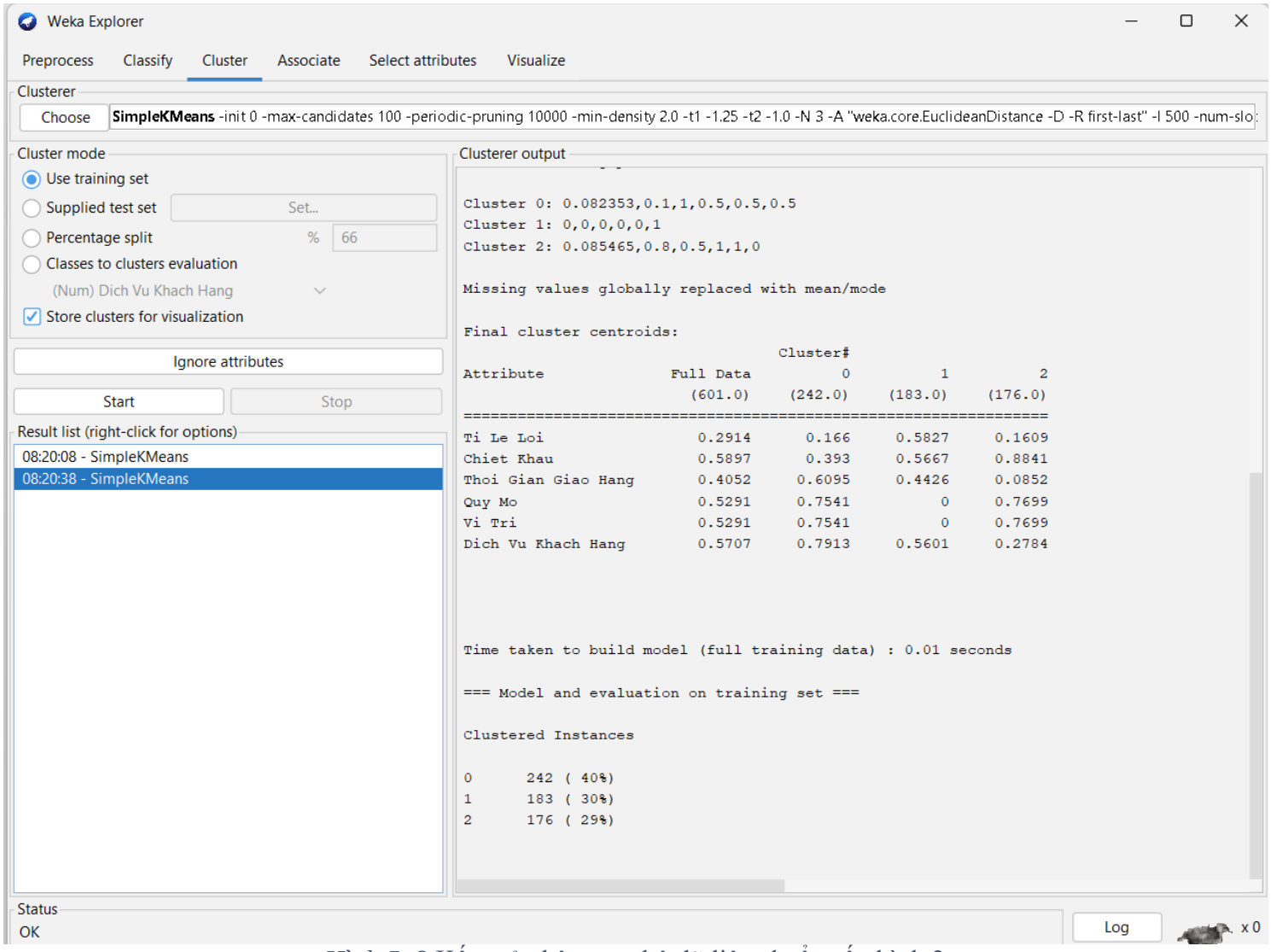
2.1.2.Cấu hình 2:

Số cụm k=3
Số vòng lặp tối đa:500
Phương pháp khởi tạo: random
Seed=40 (thay đổi seed nhằm kiểm tra độ ảnh hưởng của giá trị khởi tạo đến kết quả)



Hình 5. 7 Cấu hình 2 trong phân cụm bộ dữ liệu chuẩn

Kết quả phân cụm cấu hình 2:



Hình 5. 8 Kết quả phân cụm bộ dữ liệu chuẩn cấu hình 2

2.1.3.Đánh giá kết quả:

Thông qua quá trình phân cụm dựa trên 2 cấu hình khác nhau đều cho kết quả phân cụm tương đương nhau.Đều này chứng tỏ:

- Cấu trúc bộ dữ liệu rõ ràng,ít nhiễu,giúp thuật toán cho ra cũng 1 kết quả cho dù có sự thay đổi về giá trị khởi tạo
- Thuật toán Kmeans hoạt động ổn định khi thay đổi giá trị seed trong trường hợp này

Kết quả phân cụm:

Cụm 0: 242 điểm(chiếm 40% bộ dữ liệu)

- Giá trị trung bình Tỉ lệ sản phẩm lỗi 0.166,Chiết khấu 0.393,Thời gian giao hàng 0.6095,Quy Mo 0.7541,Vi Tri 0.7541,Dich Vu Khach Hang 0.7913

Cụm 1: 183 điểm (chiếm 30% bộ dữ liệu)

- Giá trị trung bình Tỉ lệ sản phẩm lỗi 0.5827,Chiết khấu 0.5667,Thời gian giao hàng 0.4426,Quy Mo 0,Vi Tri 0,Dich Vu Khach Hang 0.5601

Cụm 2: 176 điểm (chiếm 29% bộ dữ liệu)

- Giá trị trung bình Tỉ lệ sản phẩm lỗi 0.1609,Chiết khấu 0.8841,Thời gian giao hàng 0.0852,Quy Mo 0.7699,Vi Tri 0.699,Dich Vu Khach Hang 0.2784
-

Đặc trưng cụm:

- Cụm 0: 242 điểm(chiếm 40% bộ dữ liệu)
  - Tỉ lệ sản phẩm lỗi Khá Tốt
  - Chiết khấu nằm ở mức thấp
  - Thời gian giao hàng Trung Bình
  - Vị trí địa lí Trung Bình
  - Quy mô Trung Bình
  - Dịch vụ khách hàng ở mức Kém

- Cụm 1:218 điểm (chiếm 30% bộ dữ liệu)
  - Tỉ lệ sản phẩm lỗi Cao
  - Chiết khấu Trung Bình
  - Thời Gian Giao Hàng Trung bình
  - Vị trí địa lí Xa
  - Quy mô Nhỏ
  - Dịch vụ khách hàng ở mức Trung Bình

- Cụm 2: 176 điểm (chiếm 29% bộ dữ liệu)
  - Tỉ lệ sản phẩm lỗi ở mức Thấp
  - Chiết khấu khá cao
  - Thời Gian giao hàng Nhanh
  - Vị trí địa lí Xa
  - Quy mô lớn
  - Dịch vụ khách hàng Tốt

Kết luận:

Theo kết quả phân cụm và đặc trưng của từng cụm,có thể kết luận như sau:

- Các nhà cung cấp ở Cụm 0 và Cụm 2 là những nhà cung cấp đánh tin cậy,có thể hợp tác vì các thuộc tính đều đưa ra kết quả từ mức Trung Bình- Cao

- Các nhà cung cấp ở Cụm 1 là những nhà cung cấp không nên hợp tác,vì có thuộc tính đưa ra ở mức thấp,tê..Không nên hợp tác với những nhà cung cấp này để tránh những trường hợp xấu không mong muốn

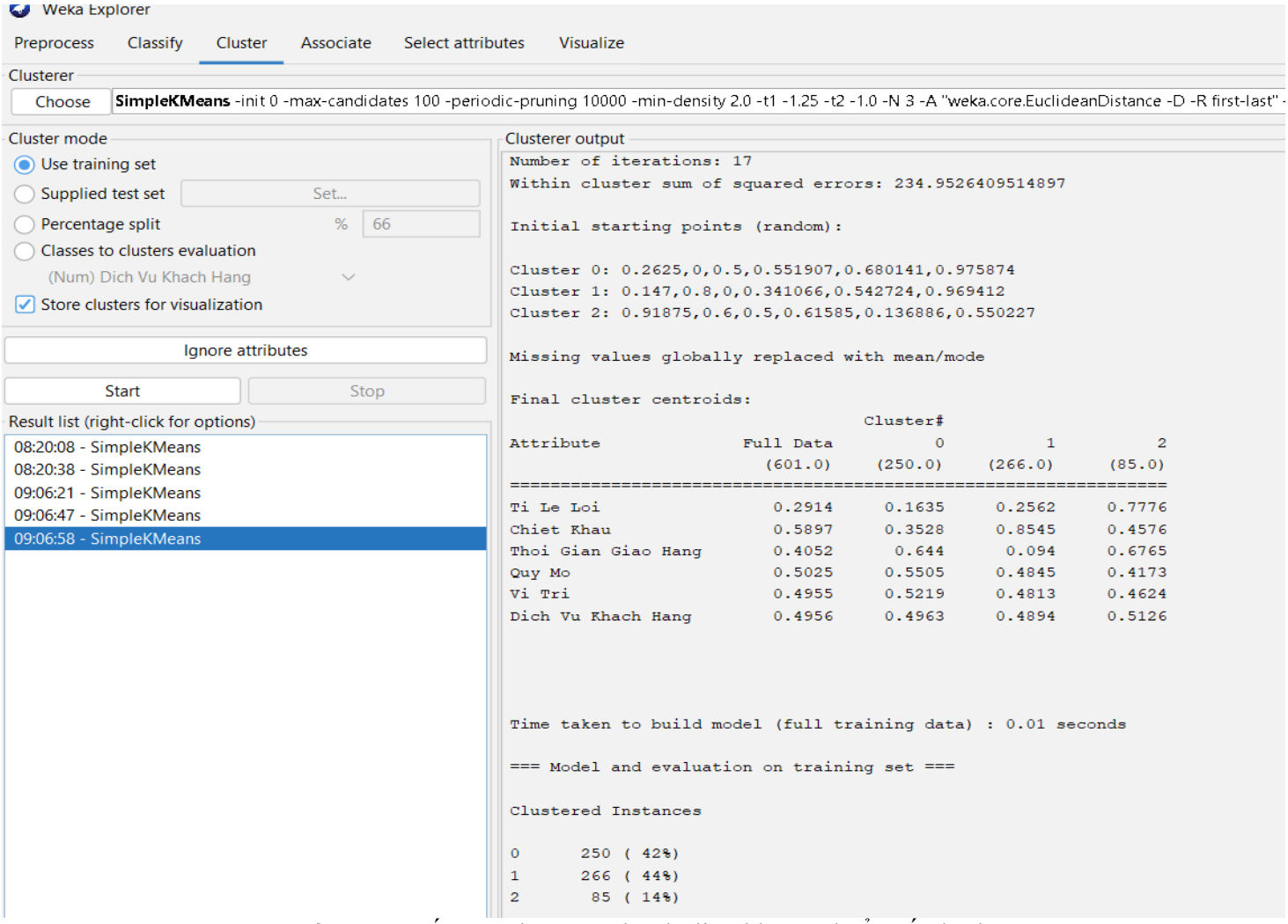
2.2.Tiến hành phân cụm bộ dữ liệu không chuẩn:

Bộ dữ liệu không chuẩn là bộ dữ liệu được tạo ra bằng cách tạo ngẫu nhiên các giá trị thuộc tính Quy Mo,Vi Tri,Dich Vu Khach Hang của bộ dữ liệu
Đặc điểm:

- Các giá trị trong bộ dữ liệu không chuẩn là giá trị ngẫu nhiên,không có quan hệ -> điều này sẽ dẫn đến kết quả gom cụm không theo quy luật

2.2.1.Cấu hình 1:

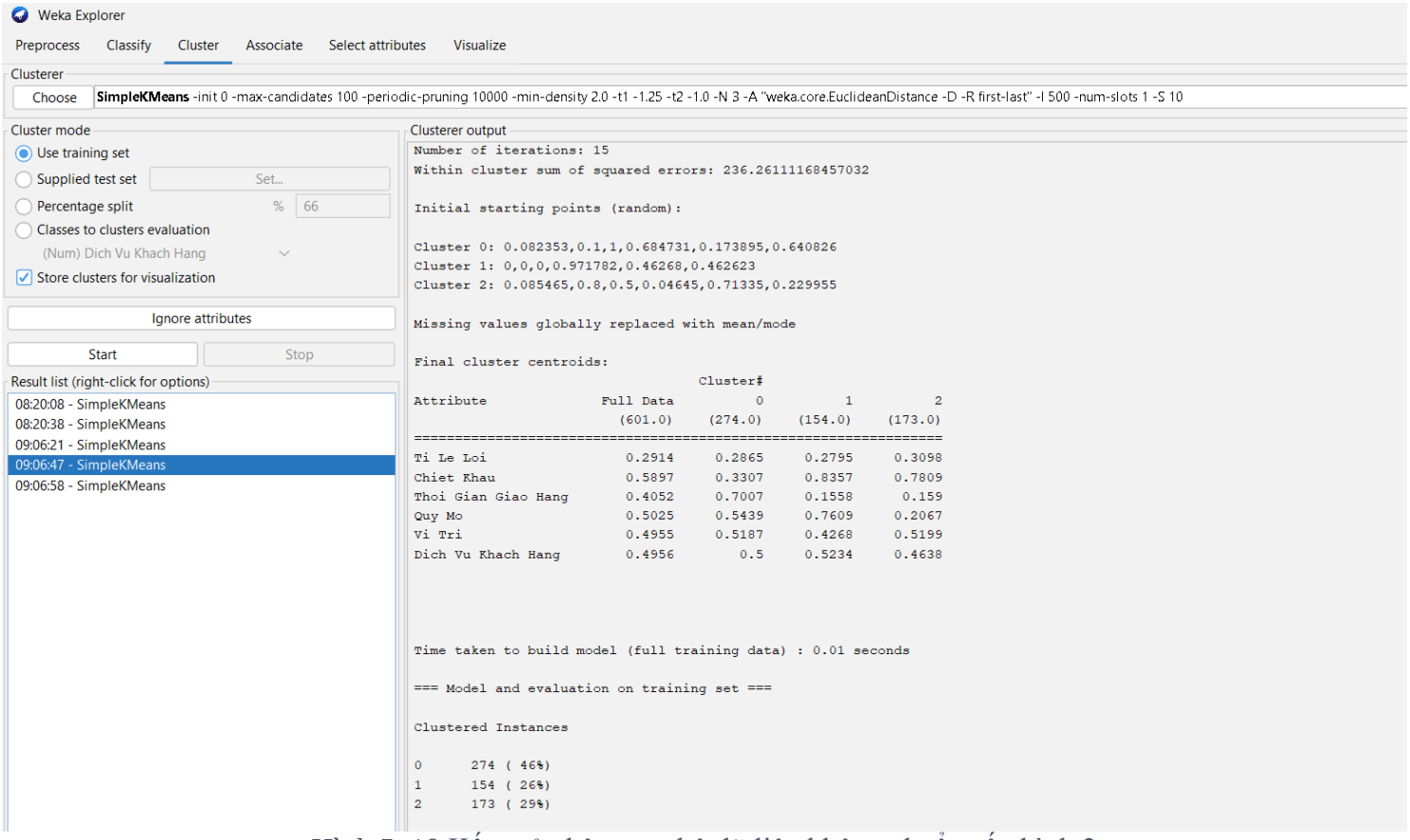
Số cụm k=3
Số vòng lặp tối đa:500
Phương pháp khởi tạo: random
Seed=10



Hình 5. 9 Kết quả phân cụm bộ dữ liệu không chuẩn cấu hình 1

2.2.2.Cấu hình 2:

Số cụm k=3
Số vòng lặp tối đa:500
Phương pháp khởi tạo: random
Seed=40



Hình 5. 10 Kết quả phân cụm bộ dữ liệu không chuẩn cấu hình 2

Đánh giá kết quả:

Thông qua quá trình phân cụm dựa trên 2 cấu hình khác nhau cho kết quả phân cụm khác nhau khá nhiều.Đều này chứng tỏ:

- Cấu trúc bộ dữ liệu không rõ ràng,các giá trị của các thuộc tính không có mối quan hệ chặt chẽ
- Các đặc trưng cụm thay đổi khi thay đổi giá trị khởi tạo.Điều này dẫn đến kết quả phân cụm bị sai lệch,không tính toán được các đặc trưng của từng cụm.

Dựa vào kết quả phân trên cho thấy: Độ chính xác bộ dữ liệu không chuẩn khá thấp,cho thấy được mức độ "Tri Thức" thấp

⇨ **Quá trình phân cụm trên bộ dữ liệu có độ chính xác khá cao,có thể áp dụng mô hình phân cụm này vào thực tế**

KẾT LUẬN

Trong quá trình đánh giá chất lượng nhà cung cấp, chúng tôi đã sử dụng nhiều tiêu chí khác nhau như tỷ lệ lỗi sản phẩm, thời gian giao hàng, quy mô, vị trí, và dịch vụ khách hàng để đánh giá và xếp hạng các nhà cung cấp. Kết quả cho thấy rằng:

- **Tỷ lệ lỗi sản phẩm**: Những nhà cung cấp có tỷ lệ lỗi thấp có xu hướng duy trì chất lượng sản phẩm ổn định và đáng tin cậy hơn.
- **Thời gian giao hàng**: Nhà cung cấp có thời gian giao hàng nhanh và đúng hẹn thường xuyên hơn được đánh giá cao, vì họ giúp duy trì sự liên tục trong hoạt động kinh doanh.
- **Quy mô và Vị trí**: Các nhà cung cấp có quy mô lớn và vị trí gần gũi với doanh nghiệp của chúng tôi thường cung cấp dịch vụ tốt hơn và linh hoạt hơn trong các tình huống khẩn cấp.
- **Dịch vụ khách hàng**: Những nhà cung cấp có dịch vụ khách hàng tốt, phản hồi nhanh chóng và giải quyết các vấn đề hiệu quả được đánh giá cao, vì họ tạo ra trải nghiệm hợp tác tích cực và giảm thiểu rủi ro cho doanh nghiệp.

Tóm lại, để lựa chọn nhà cung cấp phù hợp, doanh nghiệp cần cân nhắc tổng hợp các yếu tố trên để đảm bảo không chỉ chất lượng sản phẩm mà còn sự ổn định và hiệu quả trong quá trình hợp tác lâu dài.

Bài báo cáo Xây dựng kho dữ liệu và khai phá dữ liệu đã hoàn thành và đạt những kết quả sau:

- Tìm hiểu lý thuyết về data mining, các kĩ thuật xây dựng kho dữ liệu và khai phá dữ liệu.
- Giới thiệu Phần mềm và cách sử dụng Weka
- Xây dựng và chuẩn hoá thành công bộ dữ liệu ‘Đánh giá chất lượng nhà cung cấp’
- Giới thiệu mà sử dụng thuật toán Kmeans trong gom cụm và đánh giá bộ dữ liệu

TÀI LIỆU THAM KHẢO

1.Ứng dựng Weka

https://meeyland.com/tin-tuc/weka-la-gi-phan-mem-khai-pha-du-lieu-so-1-hien-nay-110378159473

2.Chuẩn hoá dữ liệu Min – Max:

https://viblo.asia/p/scaling-vs-normalization-oOVlYJJz58W